Large Stakes and Big Mistakes

DAN ARIELY

Duke University, Fuqua School of Business

URI GNEEZY

University of California San Diego, Rady School of Management

GEORGE LOEWENSTEIN

Carnegie Mellon University, Social and Decision Sciences

and

NINA MAZAR

University of Toronto, Rotman School of Management

Published in Review of Economic Studies (2009), 76 (2), p. 451-469.

ABSTRACT

Workers in a wide variety of jobs are paid based on performance, which is commonly seen as enhancing effort and productivity relative to non-contingent pay schemes. However, psychological research suggests that excessive rewards can in some cases result in a decline in performance. To test whether very high monetary rewards can decrease performance, we conducted a set of experiments in the US and India in which subjects worked on different tasks and received performance-contingent payments that varied in amount from small to very large relative to their typical levels of pay. With some important exceptions, very high reward levels had a detrimental effect on performance.

Acknowledgements. The authors are grateful for the help of the faculty and students at Narayanan College in Madurai, who carried out the experiment in India. Faculty: Dr. R. Srinivasan; Prof. A. Narasimhamurthy; Dr. K. Ramasamy; Dr. M. Jayakuma. Students: J. Moses Gnanakkan; P. Kalaignar; M. Ramesh; G. Selvakumar; K. Prabakara Doss. Other thanks go to Christopher Simeone, Mark Porter, and Jackie Zires for their help with running the experiments at MIT and the University of Chicago, Ricardo Paxson for designing the set of numbers for the matrices in the adding task and advising on the analyses, as well as Kristina Shampanier, On Amir, Uri Simonsohn, Avi Goldfarb, Peter Perkins, and Lowell Taylor for their invaluable insight into stats. The authors are also indebted to the editor, and the referees for their helpful comments.

1. INTRODUCTION

Payment-based performance is commonplace across many jobs in the marketplace. Many, if not most upper-management, sales force personnel, and workers in a wide variety of other jobs are rewarded for their effort based on observed measures of performance. The intuitive logic for performance-based compensation is to motivate individuals to increase their effort, and hence their output, and indeed there is some evidence that payment for performance can increase performance (Lazear, 2000).

The expectation that increasing performance-contingent incentives will improve performance rests on two subsidiary assumptions: (1) that increasing performance-contingent incentives will lead to greater motivation and effort, and (2) that this increase in motivation and effort will result in improved performance.

The first assumption, that transitory performance-based increases in pay will produce increased motivation and effort, is generally accepted, although there are some notable exceptions. Gneezy and Rustichini (2000a), for example, have documented situations, both in laboratory and field experiments, in which people who were not paid at all exerted greater effort than those who were paid a small amount (see also Gneezy and Rustichini, 2000b; Frey and Jegen, 2001; Heyman and Ariely, 2004). These results show that in some situations paying a small amount in comparison to paying nothing seems to change the perceived nature of the task, which, if the amount of pay is not substantial, may result in a decline of motivation and effort.

Another situation in which effort may not respond in the expected fashion to a change in transitory wages is when workers have an earnings target that they apply narrowly. For example, Camerer and colleagues (1997) found that New York City cab drivers quit early on days when their hourly earnings were high and worked longer hours when their earnings were low. The

authors speculated that the cab drivers may have had a daily earnings target beyond which their motivation to continue working dropped off.

Although there appear to be exceptions to the generality of the positive relationship between pay and effort, our focus in this paper is on the second assumption – that an increase in motivation and effort will result in improved performance. The experiments we report address the question of whether increased effort necessarily leads to improved performance. Providing subjects with different levels of incentives, including incentives that were very high relative to their normal income, we examine whether, across a variety of different tasks, an increase in contingent pay leads to an improvement or decline in performance. We find that in some cases, and in fact most of the cases we examined, very high incentives result in a decrease in performance. These results provide a counterexample to the assumption that an increase in motivation and effort will always result in improved performance.

2. PRIOR RESEARCH ON THE CONNECTION BETWEEN EFFORT AND PERFORMANCE

Unlike the relationship between pay and motivation/effort, the relationship between motivation/effort and performance has not attracted much attention from economists, perhaps because the belief that increased motivation improves performance is so deeply held. However, research by psychologists has documented situations in which increased motivation and effort can result in a decrement in performance – a phenomenon known as "choking under pressure" (Baumeister, 1984).

Although conventional economics assumes a positive relationship between effort and performance, there are a wide range of psychological mechanisms that could produce the

opposite relationship. These include: increased arousal, shifting mental processes from "automatic" to "controlled," narrowing of attention, and pre-occupation with the reward itself.

The increased arousal account is embodied in the "Yerkes-Dodson law" (Yerkes and Dodson, 1908), which posits that there is an optimal level of arousal for executing tasks, and that departures from this level in either direction can lead to a decrement in performance. The effect was first demonstrated by Yerkes and Dodson with rats that were placed in a cage and forced to repeatedly choose between exploring either the left or right of two passages. On each trial, the experimenters randomly hung a white card in one passage and a black card in the other. While exploring the passage with the white card resulted in a reward, exploring the passage with the black card always resulted in a shock. For some rats the shock was always small, for some medium, and for a third group it was strong. The main finding was that the rats learned to avoid the shocks most quickly when the shocks were at an intermediate level of intensity. Related results have been obtained in human motor performance (Neiss, 1988), when arousal was increased by stimulant drugs, muscle tension, or electric shocks. Since arousal is tightly linked to motivation and performance, the "Yerkes-Dodson law" implies that increases in motivation beyond an optimal level can, in some situations, produce supra-optimal levels of arousal and hence decrements in performance.

Another possible mechanism for the negative effects of increased incentives is that increased incentives can cause people, involuntarily, to consciously think about the task, shifting control of behavior from "automatic" to "controlled" mental processes even though it is well documented that controlled processes are less effective for tasks that are highly practiced and automated (Langer and Imber, 1979; Camerer et al., 2005). Sports provide a prototypical example of such over-learned, automatic tasks. Thinking about how one is swinging the golf

club or bat, or about how to get the basketball into the net can have perverse effects on performance. In fact, there are some studies of choking under pressure in sports, including one Australian study, which found that free-throw shooting performance among elite Australian basketball players was worse during games than during training (Dandy et al., 2001). This mechanism can also help to explain why the presence of an audience or competition, which tends to increase the motivation to perform well, can have detrimental effects (see also Zajonc, 1965).

A third mechanism by which increased motivation is likely to have a negative effect on performance involves the focus of attention. Increased motivation tends to narrow individuals' focus of attention on a variety of dimensions (Easterbrook, 1959), including the breadth of the solution set people consider. This can be detrimental for tasks that involve insight or creativity, since both require a kind of open-minded thinking that enables one to draw unusual connections between elements. McGraw and McCullers (1978) provided support for this mechanism by showing that the introduction of monetary rewards for tasks that involved problem-solving had detrimental effects on performance. In addition to the narrowing of attention, large incentives can simply occupy the mind and attention of the laborer with thoughts about her future in case she would get the reward and her regrets if she would not, distracting her from the task at hand.

In summary, psychological research has identified several mechanisms that can produce choking under pressure, suggesting that there are diverse factors that can create the type of pressure that produces choking. The sources most relevant for this type of pressure seem to be the presence of an audience (passive onlookers), competition (i.e. presence of others involved in the same activity; "coaction effect"), personal traits such as competitiveness, and ego-relevant threats like the belief that a task is diagnostic of something, such as intelligence, that one cares

about (see Baumeister and Showers, 1986, for an in-depth discussion of these effects; McGraw, 1978; Zajonc, 1965; also Frey and Jegen, 2001).

For economics, however, the most interesting determinant of performance pressure is the level of performance-contingent monetary incentives – an important variable for most economic behavior and labor (i.e. situations driven by external motivation). Accordingly, experiments in economics have examined the effects of the magnitude of monetary incentives on decisions. For example, Slonim and Roth (1998) report the results of an experiment with repeated ultimatum games conducted in the Slovak Republic in which financial incentives were varied by a factor of 25. Their findings suggest that increasing incentives has only a small effect on behavior of inexperienced participants, but a larger effect as participants gain experience with the game: Experienced participants rejected offers less frequently and made lower offers as the stakes increased (although see Cameron, 1999, who found no difference in behavior when stakes were changed in the ultimatum game).

In a review of the experimental literature on the effect of incentives, Camerer and Hogarth (1999) report 74 studies in which the level of pay was varied in different kinds of tasks (e.g., bargaining, trading, choosing, problem solving). They reach an inconclusive view: the majority of studies do not find any effect on performance, but many studies did observe a decrease in variance, arguably because people put more effort into the task. A few papers reported an increase in performance, and even fewer reported a decrease. Decreases were typically found in prediction tasks or tasks in which simple intuition or habit provides an optimal answer and thinking harder makes things worse (see e.g., Arkes, Dawes, and Christensen, 1986; Ashton, 1990; and Hogarth et al., 1991). Most of the studies surveyed in Camerer and Hogarth (1999) look only at the difference between no pay and low pay, which is different from our goal

in the current paper. The studies that do use high pay and compare it to low pay generally do not show an effect. However, the high pay in these papers is still far from the level of pay we use, and the tasks tend to be quite different.

Our primary goal in the studies reported herein is to test, in experiments that satisfy standard experimental economics criteria, whether increasing monetary incentives beyond some threshold may result in lower performance. A second major goal that distinguishes our work from previous contributions is to examine the generality of any detrimental effect of incentives. Among the six tasks in the first experiment, therefore, we included some that drew primarily on motor skills, some that drew primarily on memory, and some that drew primarily on creativity. Based on the literature showing detrimental effects of incentives on motor skills and creativity, we anticipated that the high monetary rewards might interfere with tasks that draw primarily on these skills, but not with those involving primarily memory. As will be seen, however, no such differences emerged; the highest levels of monetary rewards produced lower performance on all tasks in the first experiment. To examine this issue further in the second experiment, we included a task that required only physical effort. Such a task should not be subject to any of the mechanisms leading to choking under pressure as identified in the psychology literature. In this case the predicted differences between tasks did emerge. Finally, our third experiment extends the scope of investigation from financial to social incentives.

3. EXPERIMENT 1

3.1. *Design*

Eighty-seven residents of a rural town in India were recruited to participate in the experiment, which took place late in 2002.[1] Subjects were recruited by word of mouth in the village. The researchers first collected names of people interested in participating, and then contacted interested individuals to schedule experimental sessions. The sample consisted of 26.4 percent females and 73.6 percent males. The majority of participants (90.8 percent) were Hindu, 5.7 percent were Christian, and 3.4 percent were Muslim. The standard of living of our participants can be best described by their level of education and type of possessions. Participants in this experiment had, on average, 5.6 years of education, and 26.4 percent had no formal education. Approximately half of the participants reported that they owned a TV (M = 49.4 percent), and about half owned a bicycle for transportation (M = 51.7 percent). None owned a car, and only 6.9 percent had a telephone in their house.

The experiment was conducted with one participant at a time. Participants were randomly assigned to one of three treatments in which they faced incentives (on all games) that were either relatively small, moderate, or very large. In each treatment, participants played six different games in a random order and were promised payments for each game if they reached certain performance levels. The magnitude of the payment in each game depended on the treatment (low, mid or high incentive magnitude) and whether they reached either of two specified performance levels which we labeled "good" and "very good." In each game, participants received full payment (i.e. 4, 40 or 400 Indian Rupee Rs depending on the treatment) if they

[1] The experiment was conducted by local research assistants from Narayanan College at Madurai, India, who were naïve to the hypotheses.

reached the "very good" performance level, half of that if they reached the "good" performance level, and nothing if they failed to reach the "good" performance level (these two performance levels, as well as the games themselves, were selected based on pre-testing with MIT students).

The maximum possible payment for any one game in the high incentive treatment (Rs 400) was relatively close to the all-India average monthly per capita consumer expenditure (MPCE) in rural areas, which was Rs 495 (Rangachari, 2003).[2] Thus, in the unlikely event that a participant in the high payment treatment achieved "very good" performances on all six games, she would earn an amount approximately equal to half of the mean yearly consumer expenditure in the village. These stakes are effectively much larger than those that are typically offered in experimental settings.

The Games – The six games fell into three broad categories based on whether they required primarily creativity, memory, or motor skills.

The game that was used as a creativity task was "Packing Quarters." In this game participants were asked to fit 9 metal pieces of quarter circles into a black wooden frame within a given time. It is easy to fit 8 pieces, but, to fit all 9, the pieces have to be packed in a particular way. The good performance level was defined by a completion of the game within 240 seconds. The very good performance level was defined by a completion of the game within 120 seconds. Participants had only 1 trial to reach these goals.

The memory tasks included two games: "Simon" and "Recall last 3-digits." "Simon" is an electronic game that requires memory and repetitions. The game flashes a sequence of colored lights accompanied by the light-specific sounds, and the goal is to repeat the sequence by pushing the corresponding light-buttons in the same order. The good performance level was

[2] The conversion is based on the average exchange rate in December 2002 of Indian Rupee Rs 47.93 = US $1 (see Federal Reserve Statistical Release, 2003).

defined by at least one repetition of 6 consecutive lights. The very good performance level was defined by at least one repetition of 8 consecutive lights. Participants had 10 trials to reach these goals. The second memory game was "Recall last 3-digits" in which the experimenter read a sequence of digits, stopped at an unannounced point, and the participant was asked to recall the last 3-digits. Participants had 14 trials in this game. The good performance level was defined by at least 4 correct trials. The very good performance level was defined by at least 6 correct trials.

Finally, there were three different motor skill tasks: "Labyrinth", "Dart Ball", and "Roll-Up". "Labyrinth" is a game with a playing surface on top of a box that can be tilt in either of two planes. The playing surface shows a pathway from the "start" position along which the player has to advance a small steel ball to the "finish" position while avoiding the traps (holes in the board). The good performance level was defined by passing the 7^{th} hole. The very good performance level was defined by passing the 9^{th} hole. Participants had 10 trials to reach these goals. "Dart Ball" is similar to Darts but instead of throwing sharp metal arrows, the game uses tennis balls thrown at an inflated target with Velcro patches. Participants had 20 trials in this game. The good performance level was defined by throwing at least 5 balls onto the center of the target. The very good performance level was defined by throwing at least 8 balls onto the center of the target. "Roll-Up" is a game in which one attempts to drop a ball into the highest possible slot by deftly spreading apart then pushing together two rods (Baumeister, 1984). Participants had 20 trials in this game. The good performance level was defined by dropping at least 4 balls into the furthest slot. The very good performance level was defined by dropping at least 6 balls into the furthest slot.

3.2. *Results*

There are four possible ways to treat the dependent measures in this experiment: One would be to look at the raw scores, as presented in Table 1.

••• TABLE 1 •••

But raw scores do not directly relate to the compensation participants received. A second way to present the data (see Table 2) is based on the probability of reaching at least the "good" performance level at which participants received at least half pay (i.e. 2, 20, or 200 Rs.), and the probability of reaching the "very good" performance level at which participants received full pay (i.e. 4, 40, or 400 Rs.).

••• TABLE 2 •••

A final approach would be to examine the fraction of earnings from the total possible earnings (percent of maximal earnings). Since for each game participants could either earn 0 percent, 50 percent, or 100 percent of the total possible earnings this measure maps 1-to-1 onto actual performance level reached (see Table 3). The general pattern of conclusions was the same regardless of how we analyzed the data. The most interesting measure from an economics perspective is the fraction of possible earnings, since it represents the measure that is most closely linked to the incentives that the participants actually faced. In what follows, therefore, we present all results in terms of this measure.

••• TABLE 3 •••

As can be seen in Figure 1a, the aggregated performance levels across all six games (measured as the average fraction of maximum possible earnings) shows that relatively high monetary incentives can have perverse effects on performance. The average share of earnings relative to maximum possible earnings was lowest in the high payment condition (M = 19.5 percent, Std. Dev. = 30.3), but higher and almost equal in the mid (M = 36.7 percent, Std. Dev. = 40.1) and low payment conditions (M = 35.4 percent, Std. Dev. = 42.5). The results of a linear regression with robust standard errors in which the dependent measure was the performance across all six games and the independent variables were dummies for the two incentive levels mid and high are reported in Table 4. Together these findings support the main hypothesis that motivated the experiment – namely that additional incentives can decrease performance.

••• TABLE 4 •••

Somewhat contrary to our expectations, however, the pattern of results held across tasks differing both in terms of difficulty and the types of skills they require (see Figure 1 b-d). To test for the significance of observed differences, we analyzed the data separately for each of the games with an ordered probit in which the dependent measure was performance in a game (measured as fraction of maximum possible earnings) and the independent variables were dummies for two incentive levels low and mid.[3] Results are presented in Table 5.[4]

[3] The six different games might have different cut-points. Therefore, running separate specifications for each game enables different cut-points for each.

[4] We also tested models, which included socio-demographic variables and their interactions with the payment condition. In no case were the socio-demographic variables significant, and, as a consequence, they are not considered in the analyses we report.

••• FIGURE 1 •••

The comparison of the low and mid levels of incentives revealed little difference in performance: Only one of the games (Labyrinth) showed a marginally significant effect. Comparisons between the low payment condition and the high payment conditions, however, revealed a number of statistically significant differences (see Table 5). The contrasts were significant at the 0.05 level for Packing Quarters, Simon, and Labyrinth, and not significant for Recall last 3 digits, Dart-Ball, and Roll-Up.

••• TABLE 5 •••

3.3. *Summary*

Overall, the results point to three main conclusions: First, with the sole exception of the Labyrinth game there was no significant difference in the performance between the low and mid payment conditions. Thus, despite the relative large difference in magnitude of reward across the treatments (i.e. 10 times higher for the mid payment condition relative to the low payment condition), performance did not seem to increase. One interpretation of this result is that the incentives in the low payment condition (which were not altogether that low) created a level of performance that was already at a peak.

Second, and more importantly, the performance of participants was always lowest in the high payment condition when compared with the low and mid payment conditions together.

Third, and contrary to our expectations, we did not observe any obvious difference in the effect of incentives on performance for different categories of games. We included, for example, "Simon" and "Recall last 3 digits" because these games require tiresome memory, and we

thought that participants who were more motivated might be more likely to maintain high levels of memory. We did not, however, observe any such difference; both games generally displayed declining performance as a function of incentives – same as the motor skill tasks and the creativity task, although this pattern was significant at the 0.05 level only for three of the six games.

There are a number of possible reasons for why the two memory tasks showed the same pattern of results as the motor skills and creativity tasks, and not the pattern that we initially expected. One is that they involved cognitive skills, like attention, that may in fact be vulnerable to one or more of the choking-generating mechanisms discussed in the introduction. For example, sometimes memorizing information is actually easier if one doesn't pay too much attention. Another is that the incentives we chose may have simply been too high. Perhaps the memory tasks had higher threshold levels of motivation at which performance started to decline, but our choice of incentive levels in the three conditions, and particularly in the high incentive condition was sufficiently extreme to produce arousal that exceeded the optimal level for all tasks, masking any difference between them.

4. EXPERIMENT 2

4.1. *Design*

Experiment 1 was conducted in India, which enabled us to offer significant monetary incentives on a relatively modest budget. While the results suggest that very high incentives can be detrimental, this conclusion does suffer from some limitations that we sought to address in a follow-up study.

One limitation was that participants in Experiment 1 were unfamiliar with most of the games, raising the question of whether incentives would have different effects for more familiar tasks or tasks where people had had an opportunity to practice. Second, the experimental setup in Experiment 1 was based on a between-subject design, so that all participants completed all six games under the same incentive condition. Clearly, it would be interesting to see if the same participants would exhibit different levels of performance when confronted with different levels of incentives. It is also useful to examine whether the results of Experiment 1 could be attributed to inferences that subjects made about the difficulty of the task based on compensation levels that were offered. A within-subject design would alleviate concerns related to such inferences. Third, the unusual nature of the participant population in Experiment 1 raises questions about whether the results would generalize to people who are more used to conditions in an advanced capitalist country. Finally, while the experiment in India necessitated a rather simple reward structure due to the literacy level of the respondents, in this experiment we could incorporated a slightly more complex reward structure. Responding to all of these issues, Experiment 2 was conducted at MIT with twenty-four undergraduate student subjects, using tasks that were more familiar to participants, with practice trials, and using a within-subject design (in which each participant received both high and low levels of incentives).

A surprise from the first experiment was the failure to observe different effects of high incentives for different types of tasks, contrary to the rationale that guided the inclusion of the different tasks. In the second experiment, we attempted to address this issue again, by including one task that requires only physical effort and another that requires mainly cognitive skills. The two tasks were adding and key-pressing. In the adding task respondents were given a set of 20 matrixes one at a time, with 12 numbers in each matrix (see Figure 2 for a sample), and were

asked to find the two numbers in that matrix that would add to 10. In the key-pressing task respondents were asked to alternate between pressing the "v" and "n" keys on the keyboard. We used these tasks because they are based on simple elementary aspects of performance: adding two numbers and typing – tasks that are very familiar to our respondents. One other important aspect of these tasks is that, while the adding task requires cognitive effort, the key-pressing one requires only pure physical effort without any need for cognitive resources. Thus, we should be able to examine the first postulate – that high performance-contingent incentives increase pure effort and, as a consequence, improve performance that is based solely on pure effort – as well as the second postulate – that high performance-contingent incentives can decrease performance that is based on cognitive skills. Consequently, we expected an improvement in performance for the key-pressing task when the stakes were high. However, because the adding task required cognitive resources and effort, we predicted that increased incentives would lead to a decrement in performance on this task.

••• FIGURE 2 •••

For the adding task, performance was measured by the number of matrices that were solved correctly in four minutes. For the key-pressing task, performance was measured by the number of alternations in four minutes – a deliberately mind-numbingly boring task. The low incentive for the adding task was $0 if respondents solved 9 or fewer matrices, $15 if respondents solved 10 matrices, and an additional $1.50 for each additional matrix solved to a maximum of $30. The high incentive for the adding task was ten times higher ($0, $150, $300). The low incentive for the key-pressing task was $0 if respondents pressed 599 alternations or

less, $15 if respondents pressed 600 alternations, and an additional $0.10 for each additional alternation (based on pilot testing we expected the maximum to be 750 alternation, which would equal a payment of $30). The high incentive for the adding task was ten times higher ($0, $150, $300).

Finally, using a within-subject design also allowed us to examine the effect of increasing incentives at the individual choice level. Economists typically conceive of effort as a choice variable, meaning that if excessive effort worsens performance, then a worker would never choose to exert so much effort. While this assumption that the level of effort is a matter of choice might not always be the case (for example it might be difficult to regulate mental effort since arousal is the brain's way of increasing its level of effort, and arousal is not ordinarily under volitional control; see, e.g., Kahneman, 1973), to the extent that it is a choice variable, our participants might intuit to regulate their effort level.

The experiment was conducted toward the end of the semester, a time when the students are likely to have depleted their budget and be strapped for cash. When respondents first came to the lab they were given instructions for the adding task and were given four minutes to perform this task without any incentives. Next, they were given instructions for the key-pressing task and were given four minutes to perform this task without any incentives. After this initial practice with both tasks, half of the respondents were given the two tasks (in the same order) with low incentives, and the other half were given the two tasks (in the same order) with high incentives. After finishing the first set of tasks-for-pay, each respondent was given the two tasks again (in the same order) but this time for the level of incentives they had not yet experienced. That is, each participant participated three times in each of the two tasks: once for practice and twice for pay.

4.2. *Results*

In line with the analysis of Experiment 1, the main dependent variable in our analysis was, for each task, the participant's earnings as a fraction of total possible earnings for that task (percent of $30 in the low incentive condition and percent of $300 in the high incentive condition). To test for the significance of observed differences, we analyzed the data with a linear regression in which the independent variables were the incentive levels (dummy equal to 1 for high), the types of games (dummy equal to 1 for adding), the order of the two incentives (dummy equal to 1 for high-low), and all two-way interaction terms between them. The regression included random effects and robust clustered errors for participants, assuming non-independence of observations across trials due to a repeated measures design. The regression results are presented in Table 6.

••• TABLE 6 •••

Our analyses revealed a highly significant interaction between incentive level and task. As can be seen in Figure 3, the results for the adding task replicated the basic results from Experiment 1, with performance decreasing as a function of stakes [Low: M = 62.9 percent, Std. Dev. = 23.2, M = 42.9 percent, Std. Dev. = 30.7, $\chi^2(1) = 7.22$, $p = 0.0072$], while the results from the key-pressing task showed the opposite pattern: performance increasing as a function of stakes [Low: M = 39.9 percent, Std. Dev. = 37.1, High: M = 77.9 percent, Std. Dev. = 13.8, $\chi^2(2) = 23.16$, $p < 0.0001$].

••• FIGURE 3 •••

In order to test for the existence of consistent individual differences in the propensity to choke, we calculated the absolute amount of choking in both tasks for each participant (defined as share of earnings high payment – share of earnings low payment). Figure 4 shows the scatter plot. Several patterns can be seen in the figure. First, it is possible to see what has already been shown -- that there is choking in the higher earnings condition for the adding task, but the opposite pattern for the key pressing task. In the adding task, the majority of participants, that is 17 out of 24 participants (70.8 percent) performed worse with high levels of payment, three participants (12.5 percent) were not affected by the level of payment, and four participants (16.7 percent) improved their performance with high payment. In the key-pressing task, in contrast, the majority of participants, that is 19 out of 23 participants (82.6 percent; the low payment observation of one participant was missing) improved their performance with high levels of payment, one participant (4.3 percent) performed the same for high and low levels of payment, and only three participants (13 percent) decreased their performance with high payment.

Second, it is possible to examine the relationship between choking in the two tasks. Interestingly, it was not the case that the participants who choked in one of the tasks were the same ones who choked in the other; instead we found a significantly negative, moderate correlation ($r = -0.3968$, $t(22) = 4.8073$, $p < 0.0001$). For example, the two individuals who improved the most as the incentives increased in the adding task, ended up being the highest chokers in the key-pressing task. This individual level variation suggests that the factors leading to choking under pressure include not only individual characteristics, but also task-specific characteristics.

••• FIGURE 4 •••

The findings of Experiment 2 provide additional support for the main hypothesis that motivated the current work – namely that additional incentives can decrease performance. Combined with the findings from Experiment 1, these results also show that such negative returns to incentives can appear in tasks that respondents are generally familiar with (adding numbers), and even when they had some practice with the specific task. Furthermore, the results from Experiment 2 show that the order of the two incentive levels did not have a significant influence – suggesting that the effects are not due to inferences respondents draw about the difficulty of the task based on the level of reward. In addition, the decreased performance with high incentives observed for the adding task and the increased performance with high incentives observed for the key-pressing task support the idea that tasks that involve only physical effort are likely to benefit from increased incentives, while for tasks that include a cognitive component, such as adding numbers, there seems to be a level of incentive beyond which further increases can have detrimental effects on performance. Finally, based on the lack of a positive correlation between choking on the two tasks, this study does not provide support for the idea that there are meaningful individual differences in individuals' propensity to choke. If there are such differences, they may be task-specific.

5. EXPERIMENT 3

5.1. *Design*

Experiments 1 and 2 demonstrated that large contingent financial incentives can sometimes decrease performance. In Experiment 3 we extend the scope of the investigation to examine social incentives. An extensive literature on audience and coaction effects has shown that the

presence of passive onlookers or spectators as well as others, who engage in the same activity can significantly influence people's performance. The basic theory propagated by Zajonc (1965) in his "social facilitation" framework is that audience or coaction increases arousal, which in turns facilitates an individual's dominant response to a situation. Based on this "drive theory", audience or coaction can be positive or negative, dependent on the specific task and an individual's experience with the task. For a well-learned task, the theory predicts that the presence of others should increase performance (since with experience, the dominant response tends to produce good performance), while performance on a novel task should detriment (since the dominant response tends to result in poor performance). Follow up papers have argued for a more cognitive model, where audience and coaction effects depend less on the objective reality of a situation and more on individuals' perceptions of the situation and their personality (e.g., will others evaluate me?; will I be punished or rewarded?; do I care?), (Ferris et al., 1978; see also Baumeister, 1984; Baumeister and Showers, 1986). In an economic setting, Charness, Rigotti, and Rustichini (2007) show that peoples' behavior in games is affected by audience.

Our experiment contributes to the stream of research on audience and coaction effects in two respects. First, we further examine social incentives in the context of financial incentives. Second, we investigate possible gender differences. Specifically, we examine the impact on performance when an audience watches the subject work on a cognitive task that involves performance-contingent payment. Although audience effects might seem at first glance to be non-economic in nature, there are many tasks of great economic significance that are performed under conditions of public scrutiny. Determining whether the increased motivation brought by an audience improves or worsens performance in the context of performance-contingent payments, therefore, not only provides more basic evidence on the relationship between effort and

performance, but could also have ramifications in applied settings. In addition, prior results by Gneezy, Niederle, and Rustichini (2003) suggest that men are much more responsive to competitive incentives than women, raising the question whether there might be a gender difference in the tendency to choke under these conditions.

The experiment took about 30 minutes and was conducted in five sessions at the University of Chicago. Four of the sessions had eight participants, and one session had seven participants. Upon arriving, participants received instructions in which they were told that they would be taking part in an experiment on problem solving, and that the task in the experiment was to solve anagrams. It was explained that anagrams are jumbled letters that can be made into one, and only one, very common word. Following the instructions participants had a one-minute trial in which they were asked to solve three examples of anagrams. At the end of the practice trial the correct answers were revealed.

The experiment was based on 26 trials (10 private trials and 16 public trials) for the sessions of eight participants and on 24 trials (10 private trials and 14 public trials) for the session of seven participants, each consisting of one minute to solve three anagrams. The important feature of the design was that in the private trials all participants worked without being observed by anyone, while in the public trials, one participant chosen at random worked in plain sight of the other participants. In the public trials, a random number was drawn and the corresponding participant stood next to the experimenter and attempted to solve the anagrams in front of the entire group, using a larger version of the same page that was used when anagrams were solved in private. As a consequence, each participant participated in 10 private trials and 2 public trials.

The sequence of trials alternated between two private trials (in which everyone solved two sets of three anagrams), and four public trials (in which four different participants got up one at a time and each solved one set of three anagrams). Payment was 33 cents for every anagram successfully solved, whether in a private or public round. In addition, each participant received a flat $5 for showing up.

5.2. *Results*

The main interest in this experiment is the number of solved anagrams across the two conditions (public and private) and gender. Because the anagram task involves creativity, and because we thought that solving the anagrams in front of others would produce high levels of motivation, we predicted that the public condition (as opposed to the private condition) would lead to choking under pressure.

We first collapsed our data to get the participant's earnings as a fraction of total possible earnings in the public and private trials (percent of $9.90, i.e. 10 * 3 * $0.33, in the private condition and percent of $1.98, i.e. 2 * 3 * $0.33, in the public condition), creating two observations per participant. We analyzed the data with linear regressions in which the dependent variable was the participant's earnings as a fraction of total possible earnings, and the independent variables were the trial type (dummy equal to 1 for public), gender (dummy equal to 1 for male), and (for the full model) the interaction term between them. The regressions included random effects and robust clustered errors for participants, assuming non-independence of observations across trials due to a repeated measures design. The regression results (with and without the interaction term) are presented in Table 7.

••• TABLE 7 •••

The frequency distributions are depicted in Figure 5. Our analyses revealed a highly significant main effect for the type of trial, with higher average performance in the private condition (M = 38.5 percent, Std. Dev. = 18.5) than in the public condition (M = 22.2 percent, Std. Dev. = 20.7). This result is particularly interesting given opposite findings by Falk and Ichino (2006) under non-variable payment schemes. The difference between the two findings suggests an interaction between audience and type of payment that is worth further investigating in the future. There was, however, no evidence of any gender difference in ability to solve anagrams, nor any evidence for the two genders to be differentially influenced by the social pressure. The performance per trial was 38.9 percent (Std. Dev. = 21.3) for men and 38.3 percent (Std. Dev. = 16.4) for women in the private condition, and 21.3 percent (Std. Dev. = 19.6) for men and 23 percent (Std. Dev. = 22) for women in the public condition. Future research should examine the conditions under which gender effects are likely to arise (see also Blascovich et al., 1999).

••• FIGURE 5 •••

6. GENERAL DISCUSSION

Many institutions provide very large incentives for tasks that require creativity, problem solving, and memory. Our results challenge the assumption that increases in motivation would necessarily lead to improvements in performance. Across multiple tasks (with one important exception), higher monetary incentives led to worse performance.

The finding that performance is superior for moderate incentives relative to very high incentives is consistent with the "Yerkes-Dodson law" (Yerkes and Dodson, 1908), according to which, beyond an optimal level of arousal for executing tasks, further increases in arousal can lead to a decrement in performance. The positioning of the optimal level of arousal is likely to vary based on the task, the individual's personality, and the individual's experience with the task. In general, the optimal level of arousal should be higher for more practiced tasks, particularly if prior practice has occurred under conditions of high incentives.

Our results do not, however, provide support for the idea that there are systematic individual differences in the propensity to choke. Such differences may well exist, but the only evidence relevant to the question -- the correlation across the two tasks in study 2 between the difference in performance between the low and high stakes conditions – revealed a negative rather than a positive correlation. It is possible, however, that this negative correlation may have resulted from the fact that higher effort helped performance on the key press task but hurt performance on the adding task. If there were differences between subjects in the level of effort produced by high stakes, then those who were more motivated by high stakes may have performed better on the key press task but worse on the addition task, resulting in the observed negative correlation.

Our results also point to a new justification for the use of agents. In the standard economic analysis of the principal agent problem (e.g., Hart and Holmstrom, 1987), principals are assumed to contract with agents because they confer efficiencies, either due to skill and expertise, or a lower opportunity cost of time or effort. In Fershtman and Judd (1987) agents are used to shape the incentives in competition. More recently, Hamman, Loewenstein, and Weber (2007) have proposed that agents can also be hired to avoid moral responsibility – to do the

principal's 'dirty work'. These results suggest that an over-motivated principal might hire an agent to perform a task at a more optimal, reduced level of incentives. Although our results suggest that this might in some cases be beneficial, it requires principals to be aware of the performance-debilitating effects of high incentives, which seems unlikely. In fact, in another study not reported in detail in this paper, we gave 60 participants all the information about Experiment 1 and asked them to predict the results of the Simon and Packing Quarters games. The predictions of the respondents indicated that they expected performance to be positively and monotonically linked to level of contingent reward.

These results have important implications for research in behavioral economics. The fact that some of our tasks revealed non-monotonic relationships between effort and performance of the exact type predicted by the "Yerkes-Dodson law" cautions against generalizing results obtained with one level of incentives to levels of financial incentives that are radically different (see, e.g., Parco et al., 2002). For many tasks, introducing incentives where there previously were none or raising small incentives on the margin is likely to have a positive impact on performance. This could be true even when the level of incentives is high (Ehrenberg and Bognanno, 1990; Lazear, 2000). Our experiment suggests, however, that one cannot assume that introducing or raising incentives always improves performance. Beyond some threshold level, it appears, raising incentives may increase motivation to supra-optimal levels and result in perverse effects on performance. Given that incentives are generally costly for those providing them, raising contingent incentives beyond a certain point may be a losing proposition. Perhaps there is good reason why so many workers continue to be paid on a straight salary basis.

Our results also have implications for the debate between proponents and opponents of behavioral economics. One of the common criticisms of behavioral economics is that observed

anomalies are unlikely to occur when the stakes are high (Thaler, 1986). Although people's performance undoubtedly improves in some situations as the stakes increase, the results of the experiments reported here suggest, at a minimum, that high-payments cannot be relied upon to produce optimal behavior.

In closing, we note that academics do not seem to be immune to the effects we discuss. How many of us have found ourselves in front of an audience at a loss for words, or worse, unable to deliver due to "dry mouth" at exactly the times when it is most important to perform at our best. Indeed, one of the authors was present at what for all intents and purposes appear to be an *audition* for the most important award made in economics,[5] and observed a string of superb public speakers give presentations that were very far from the best of their careers. Certainly, there are individual differences in the impact of incentives on effort, and the impact of effort on performance. However, for some fraction of the population in some situations, increasing incentives does not seem to result in enhanced performance.

[5] Indeed, two presenters at the meeting went on to win the coveted prize a year later.

REFERENCES

ARKES, HAL R., DAWES, ROBYN M. and CHRISTENSEN, CARYN (1986), "Factor influencing the use of a decision rule in a probabilistic task", *Organizational Behavior and Human Decision Process*, **37**, 93-110.

ASHTON, ROBERT H. (1990), "Pressure and performance in accounting decision setting: Paradoxical effects of incentives, feedback, and justification", *Journal of Accounting Research*, **28**, 148-180.

BAUMEISTER, ROY F. (1984)**,** "Choking under Pressure: Self-Consciousness and Paradoxical Effects of Incentives on Skillful Performance", *Journal of Personality and Social Psychology*, **46 (3)**, 610-620.

BAUMEISTER, ROY F. AND SHOWERS, CAROLIN J. (1986), "A Review of Paradoxical Performance Effects: Choking under Pressure in Sports and Mental Tests", *European Journal of Social Psychology*, **16 (4)**, 361-383.

BLASCOVICH, JIM, MENDES, WENDY BERRY, HUNTER, SARAH B., AND SALOMON, KRISTEN (1999), "Social "Facilitation" as Challenge and Threat", *Journal of Personality and Social Psychology*, **77 (1)**, 68-77.

CAMERER, COLIN F., BABCOCK, LINDA, LOEWENSTEIN, GEORGE, AND THALER, RICHARD H. (1997), "Labor Supply of New York City Cab Drivers: One Day at a Time", *The Quarterly Journal of Economics*, **112 (2)**, 407-441.

CAMERER COLIN F. AND HOGARTH, ROBIN (1999), "The effects of financial incentives in experiments: A review and capital-labor-production framework", *Journal of Risk and Uncertainty*, **19 (1)**, 7-42.

CAMERER, COLIN F., LOEWENSTEIN, GEORGE, AND PRELEC, DRAZEN **(2005).** "Neuroeconomics: How Neuroscience Can Inform Economics", *Journal of Economic Literature*, **43 (1)**, 9-64.

CAMERON, LISA (1999), "Raising the stakes in the ultimatum game: Experimental evidence from Indonesia", *Economic Inquiry*, **37 (1)**, 47-59.

CHARNESS, GARY, RIGOTTI, LUCA, AND RUSTICHINI, ALDO. (2007), "Individual Behavior and Group Membership", *American Economic Review*, **97 (4)**, 1340-1352.

DANDY, JUSTINE, BREWER, NEIL, AND TOTTMAN, ROBIN (2001), "Self-Consciousness and Performance Decrements within a Sporting Context", *Journal of Social Psychology*, **141 (1)**, 150-152.

EASTERBROOK, JAMES A. (1959), "The Effect of Emotion on Cue Utilization and the Organization of Behavior", *Psychological Review*, **66 (3)**, 183-201.

EHRENBERG, RONALD G. AND BOGNANNO, MICHAEL L. (1990), "Do Tournaments Have Incentive Effects?", *Journal of Political Economy*, **98 (6)**, 1307-1323.

FALK, ARMIN AND ICHINO, ANDREA (2006), "Clean Evidence on Peer Effect", *Journal of Labor Economics*, **24 (1)**, 39-58.

FEDERAL RESERVE STATISTICAL RELEASE (2002), "Foreign Exchange Rates December", Website: http://www.federalreserve.gov/Releases/G5/20030102/, accessed: January 2, 2003,

FERSHTMAN, CHAIM AND JUDD, KENNETH L. (1987), "Equilibrium Incentives in Oligopoly", *American Economic Review*, **77 (5)**, 927-940.

FERRIS, GERALD R., BEEHR, TERRY A., AND GILMORE DAVID C. (1978), "Social Facilitation: A Review and Alternative Conceptual Model", *Academy of Management Review*, **3 (2)**, 338-347.

FREY, BRUNO S. AND JEGEN, RETO (2001), "Motivation Crowding Theory", *Journal of Economic Surveys*, 2001, **15 (5)**, 589-611.

GNEEZY, URI, NIEDERLE, MURIEL, AND RUSTICHINI, ALDO (2003), "Performance in Competitive Environments: Gender Differences", *The Quarterly Journal of Economics*, **118 (3)**, 1049-1074.

GNEEZY, URI AND RUSTICHINI, ALDO (2000a), "Pay Enough or Don't Pay At All", *The Quarterly Journal of Economics*, **115 (3)**, 791-810.

_____ (2000b), "A Fine is a Price", *Journal of Legal Studies*, **29 (1)**, 1-18.

HAMMAN, JOHN, LOEWENSTEIN, GEORGE, AND WEBER, ROBERTO (2007), "Self-Interest through Agency: An alternative rationale for the principal-agent relationship", Working Paper, Center for Behavioral Decision Research, Carnegie Mellon University.

HART, OLIVER D. AND HOLMSTROM, BENGT (1987), "The theory of contracts", in Bewley, T. F. (ed.), *Advances in Economic Theory, Fifth World Congress,* New York: Cambridge University Press, 71-155.

HEYMAN, JAMES AND ARIELY, DAN (2004), "Effort for Payment: A Tale of Two Markets", *Psychological Science*, **15 (11)**, 787-793.

HOGARTH, ROBIN, GIBBS, BRIAN J., MCKENZIE, CRAIG R. M., AND MARQUIS, MARGARET A. (1991), "Learning from feedback: Exactingness and incentives", *Journal of Experimental Psychology: Learning, memory and Cognition*, **17**, 734-752.

KAHNEMAN, DANIEL (1973), *Attention and Effort*, Englewood Cliffs, NJ: Prentice-Hall.

LANGER, ELLEN J. AND IMBER, LOIS G. (1979),"When Practice Makes Imperfect: The Debilitating Effects of Overlearning", *Journal of Personality and Social Psychology*, **37 (11)**, 2014-2024.

LAZEAR, EDWARD P. (2000), "Performance Pay and Productivity", *American Economic Review*, **90 (5)**, 1346-1361.

MCGRAW, KENNETH O. (1978), "The Detrimental Effects of Reward on Performance: A Literature Review and a Prediction Model", in Lepper, Mark R. and Greeene, David (eds), *The Hidden Costs of Reward: New Perspectives of Human Behaviour,* New York: Erlbaum, 33-650.

MCGRAW, KENNETH O. AND MCCULLERS, JOHN C. (1979), "Evidence of a Detrimental Effect of Extrinsic Incentives on Breaking a Mental Set", *Journal of Experimental Social Psychology*, **15 (3)**, 285-294.

NEISS, ROB (1988), "Reconceptualizing Arousal: Psychological States in Motor Performance", *Psychological Bulletin*, **103 (3)**, 345-366.

PARCO, JAMES E. RAPOPORT, AMNON, AND STEIN, WILLIAM E. (200**2),** "Effects of Financial Incentives on the Breakdown of Mutual Trust", *Psychological Science*, **13 (3)**, 292-297.

RANGACHARI, DILIP (2003**),** "Poverty Down But Urban-Rural Divide Sharp", *The Times of India*, March 20, 2003, Website: http://timesofindia.indiatimes.com/articleshow/-40894515.cms.

SLONIM, ROBERT AND ROTH, ALVIN E. (1998**),** "Learning in high stakes ultimatum games: An experiment in the Slovak Republic", *Econometrica*, **66 (3)**, 569-596.

THALER, RICHARD H. (1986), “The Psychology and Economics Conference Handbook-Comment”, *Journal of Business,* **59 (4)**, S279-S284.

YERKES, ROBERT M. AND DODSON, JOHN D. (1908), ”The Relationship of Strength of Stimulus to Rapidity of Habit-Formation”, *Journal of Comparative Neurology of Psychology*, **18 (5),** 459-482.

ZAJONC, ROBERT B. (1965), “Social Facilitation”, *Science*, **149 (3681)**, 269-274.

TABLE 1

Raw Scores by Game and Treatment

	Mean raw score (Std. Dev.)		
	Low	Mid	High
Packing Quarters	202.0 (65.4)	185.7 (70.5)	235.9 (12.9)
Simon	6.5 (2.1)	6.3 (1.4)	5.2 (1.4)
Recall last 3-digits	4.9 (2.7)	5.5 (2.8)	4.6 (2.4)
Labyrinth	5.9 (2.4)	4.6 (1.8)	4.1 (1.8)
Dart Ball	2.8 (2.0)	3.6 (2.6)	2.9 (1.7)
Roll-Up	1.8 (2.1)	1.8 (3.1)	1.2 (1.5)

Notes: The interpretation of raw scores differs across games, and is as follows: (i) Packing Quarters, task completion time in seconds; completion times above 240 seconds were coded as 240 seconds; (ii) Simon, best trial: number of consecutive lights; (iii) Recall last 3-digits, number of correct trials out of 14; (iv) Labyrinth, greatest number of holes passed in 10 trials; (v) Dart Ball, number of balls hitting center in 20 trials; (vi) Roll up, number of balls hitting the furthest slot in 20 trials. For all tasks except for Packing Quarters higher raw scores indicate better performances. There were 28 observations in the low treatment, 30 observations in the mid treatment, and 29 observations in the high treatment.

TABLE 2

Success by Game and Treatment

	Percent at least "good"			Percent "very good"		
	Low (2 Rs.)	Mid (20 Rs.)	High (200 Rs.)	Low (4 Rs.)	Mid (40 Rs.)	High (400 Rs.)
Packing Quarters	28.6	43.3	10.3	25.0	33.3	0
Simon	64.3	76.7	44.8	32.1	16.7	3.4
Recall last 3-digits	64.3	73.3	58.6	42.9	36.7	20.7
Labyrinth	64.3	50.0	27.6	21.4	3.3	3.4
Dart Ball	25.0	40.0	37.9	10.7	23.3	6.9
Roll-Up	25.0	23.3	17.2	21.4	20.0	3.4

Notes: The table shows the percent of individuals who reached at least the "good" performance level and the percent of individuals who reached the "very good" performance level. There were 28 observations in the low treatment, 30 observations in the mid treatment, and 29 observations in the high treatment.

TABLE 3

Percent of Maximal Earnings by Game and Treatment

	Mean percent of maximal earnings (Std. Dev.)		
	Low	Mid	High
Packing Quarters	26.8 (44.1)	38.3 (46.8)	5.2 (15.5)
Simon	48.2 (41.9)	46.7 (32)	24.1 (28.7)
Recall last 3-digits	53.6 (45)	55 (40.2)	39.7 (38.7)
Labyrinth	42.9 (37.8)	26.7 (28.6)	15.5 (27.1)
Dart Ball	17.9 (33.9)	31.7 (42.5)	22.4 (31.6)
Roll-Up	23.2 (41.9)	21.7 (40.9)	10.3 (24.6)

Notes: Due to the ordinal nature of earnings (low: 0/2/4, mid: 0/20/40, high: 0/200/400) in this experiment, "mean percent of maximal earnings" can also be calculated from the data in table 2 as [(percent at least "good" + percent "very good")/2]. There were 28 observations in the low treatment, 30 observations in the mid treatment, and 29 observations in the high treatment.

TABLE 4

Linear Regression Results of Experiment 1 in India

	Dummies	Coef. (Robust Std. Err.) All Six Games
Payment	Mid	0.0120 (0.0449)
	High	-0.1598** (0.0414)
Constant		0.3546** (0.0321)
Observations		87
R^2		0.1962

Notes: The average performance across all six games (measured as mean fraction of maximum possible earnings) represents the dependent measure and the independent variables are dummies for the incentive levels. The linear regression includes robust standard errors. Significant differences (p <= 0.1) are marked +, (p <= 0.05) are marked *, (p <= 0.01) are marked **.

TABLE 5

Ordered Probit Results of Experiment 1 in India

	Dummies	Coef. (Robust Std. Err.)					
		Packing Quarters	Simon	Recall	Labyrinth	Dart Ball	Roll-Up
Payment	Mid	0.3456 (0.3443)	-0.0459 (0.3082)	0.0346 (0.3099)	-0.5424+ (0.2958)	0.4709 (0.3488)	-0.0539 (0.3701)
	High	-0.8622* (0.3713)	-0.8037* (0.3234)	-0.3827 (0.3101)	-1.03** (0.3433)	0.2204 (0.3267)	-0.4234 (0.3542)
Observations		87	87	87	87	87	87
Log likelihood		-60.2965	-85.3149	-94.315	-75.214	-75.1984	-56.7418

Notes: The performance in the games (measured as fraction of maximum possible earnings) represents the dependent measure and the independent variables are dummies for the two incentive levels mid and high. The ordered probit analyses include robust standard errors. Significant differences ($p <= 0.1$) are marked +, ($p <= 0.05$) are marked *, ($p <= 0.01$) are marked **.

TABLE 6

Linear Regression Results of Experiment 2 at MIT

	Coef. (Robust Clustered Err.)
High Payment	0.3419**
	(0.0817)
Adding Task	0.1254
	(0.1088)
High Payment First	-0.1823
	(0.1113)
High Payment x Adding Task	-0.5638**
	(0.1277)
High Payment x High Payment First	0.0477
	(0.0821)
Adding Task x High Payment First	0.2473*
	(0.1)
Constant	0.474**
	(0.0935)
Observations	95
Groups	24
R^2	0.2543

Notes: We analyze the data with a linear regression in which the dependent variable for each task represents the participant's earnings as a fraction of total possible earnings for that task (percent of $30 in the low incentive condition and percent of $300 in the high incentive condition). The independent variables are the incentive levels (dummy equal to 1 for high), the types of games (dummy equal to 1 for adding), the order of the two incentives (dummy equal to 1 for high-low), and all the two-way interaction terms between them. The regression includes random effects and robust clustered errors for participants, assuming non-independence of observations across trials due to a repeated measures design. One key-pressing observation with low payment is missing. Significant differences p <= 0.1 are marked +, p <= 0.05 are marked *, p <= 0.01 are marked **.

TABLE 7

Linear Regression Results of Experiment 3 at the University of Chicago

	(1) Direct Effects Only	(2) Full Model
	Coef. (Robust Clustered Err.)	Coef. (Robust Clustered Err.)
Public	-0.1633**	-0.1524**
	(0.0352)	(0.0468)
Male	-0.0054	0.0064
	(0.0535)	(0.0619)
Public x Male		-0.0236
		(0.0712)
Constant	0.388**	0.3825**
	(0.0359)	(0.0361)
Observations	78	78
Groups	39	39
R^2	0.1506	0.1513

Notes: We first collapse our data to get the participant's earnings as a fraction of total possible earnings in the public and private trials, thus there are two observations per participant. We analyze the data with linear regressions in which the dependent variable is the participant's earnings as a fraction of total possible earnings, and the independent variables are the trial type (dummy equal to 1 for public), gender (dummy equal to 1 for male), and (for the full model) the interaction term between them. The regressions include random effects and robust clustered errors for participants, assuming non-independence of observations across trials due to a repeated measures design. Significant differences p <= 0.1 are marked +, p <= 0.05 are marked *, p <= 0.01 are marked **.

FIGURE CAPTIONS

FIGURE 1: Means of the Share of Earnings Relative to the Maximum Possible Earnings for the Three Payment Levels. For All Six Games Combined a, and Plotted Separately by Game b-d. Games are Indicated by their Category: Motor Skills (ms), Memory (mm), and Creativity (cr).

FIGURE 2: Sample Screen with Matrix in Adding Task.

FIGURE 3: Means of the Share of Earnings Relative to the Maximum Possible Earnings for Key-Pressing and Adding.

FIGURE 4: Participants' Absolute Amount of Choking by Task.

FIGURE 5: Frequency Distribution of Share of Earnings Relative to the Maximum Possible Earnings For the Public and Private Conditions.

FIGURE 1

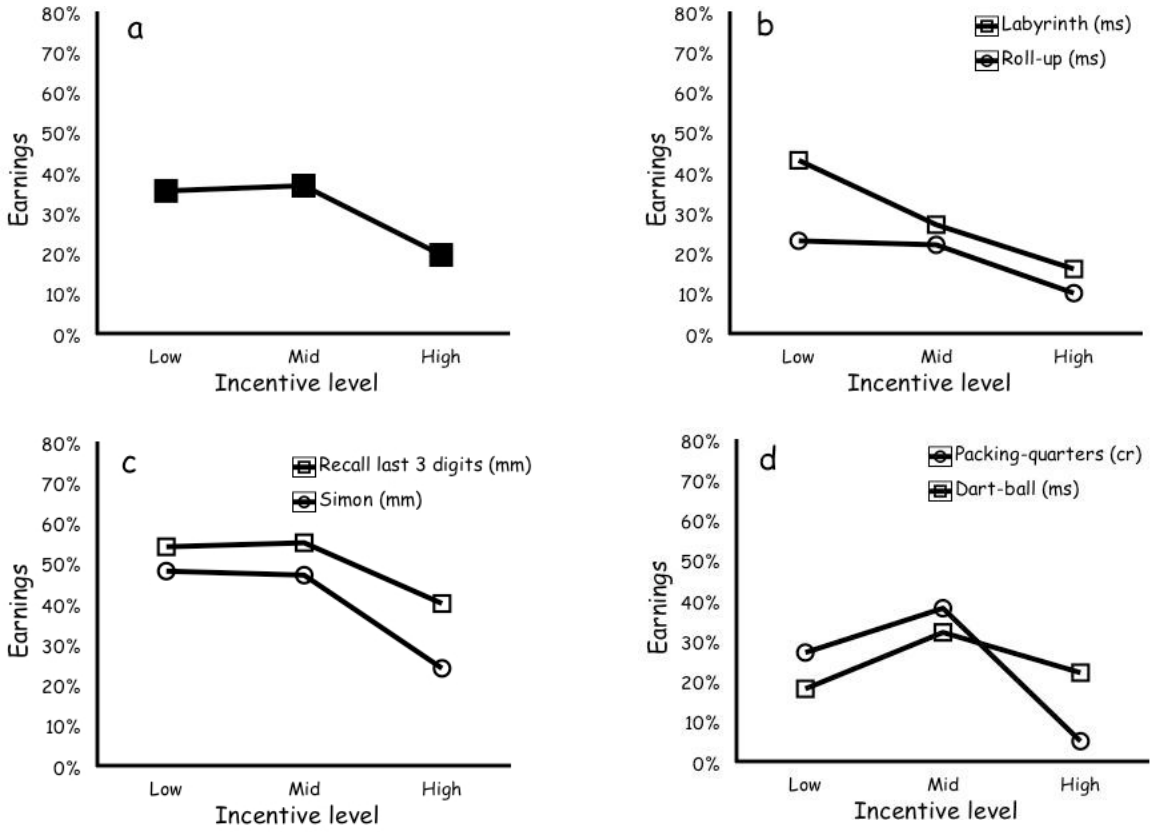

a
b
c
d
Earnings
Incentive level
Low
Mid
High
80%
70%
60%
50%
40%
30%
20%
10%
0%
Labyrinth (ms)
Roll-up (ms)
Recall last 3 digits (mm)
Simon (mm)
Packing-quarters (cr)
Dart-ball (ms)

FIGURE 2

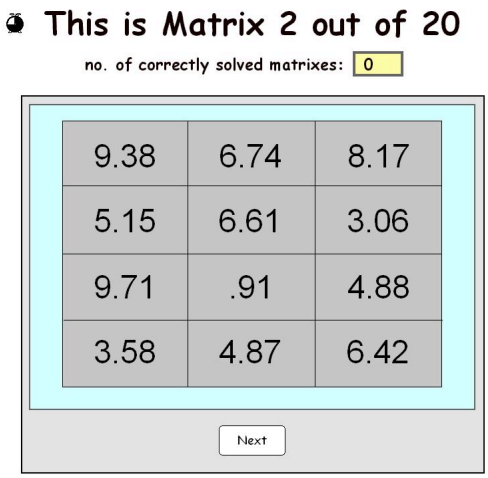

This is Matrix 2 out of 20
no. of correctly solved matrixes: 0
9.38
6.74
8.17
5.15
6.61
3.06
9.71
.91
4.88
3.58
4.87
6.42
Next

FIGURE 3

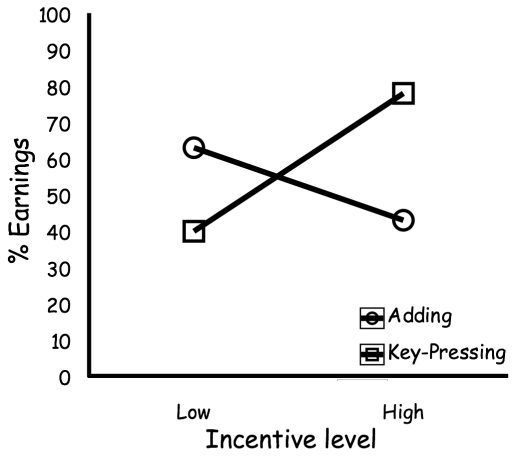

100
90
80
70
60
50
40
30
20
10
0
% Earnings
Adding
Key-Pressing
Low
High
Incentive level

FIGURE 4

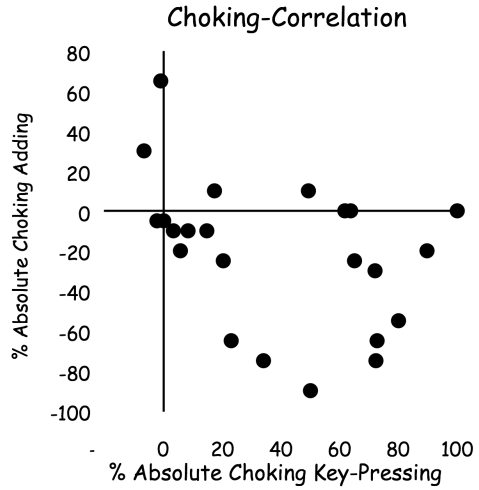
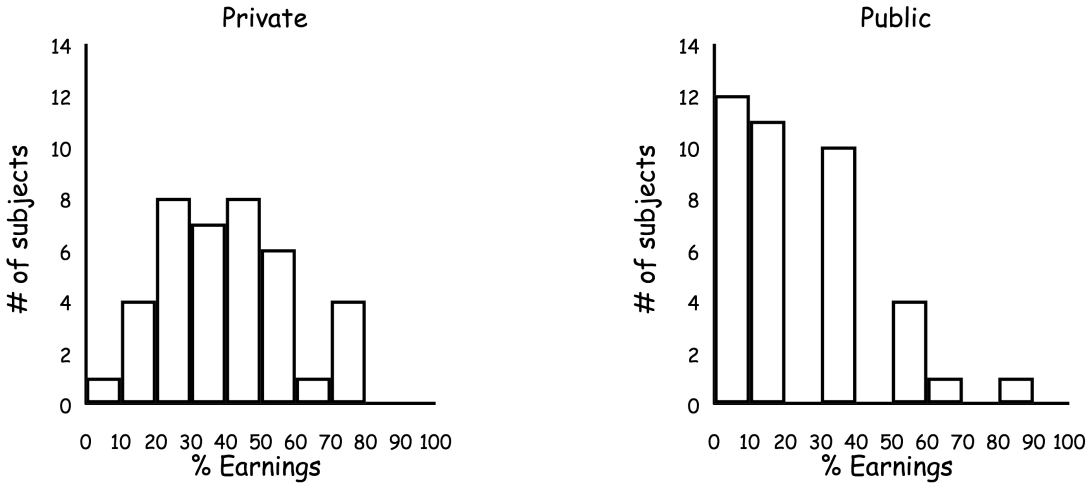

Choking-Correlation
% Absolute Choking Adding
80
60
40
20
0
-20
-40
-60
-80
-100
-
0
20
40
60
80
100
% Absolute Choking Key-Pressing

FIGURE 5


Private
# of subjects
14
12
10
8
6
4
2
0
0 10 20 30 40 50 60 70 80 90 100
% Earnings
Public
# of subjects
14
12
10
8
6
4
2
0
0 10 20 30 40 50 60 70 80 90 100
% Earnings